# An Empirical Study of Black-box based Membership Inference Attacks on a Real-World Dataset

Yujeong Kwon, Simon S. Woo, and Hyungjoon Koo*

Sungkyunkwan University, Suwon, South Korea
{shr2008, swoo, kevin.koo}@g.skku.edu

**Abstract.** The recent advancements in artificial intelligence drive the widespread adoption of Machine-Learning-as-a-Service platforms, which offers valuable services. However, these pervasive utilities in the cloud environment unavoidably encounter security and privacy issues. In particular, a membership inference attack (MIA) poses a threat by recognizing the presence of a data sample in a training set for the victim model. Although prior MIA approaches underline privacy risks repeatedly by demonstrating experimental results with standard benchmark datasets such as MNIST and CIFAR. However, the effectiveness of such techniques on a real-world dataset remains questionable. We are the first to perform an in-depth empirical study on black-box based MIAs that hold realistic assumptions, including six metric-based and three classifier-based MIAs with the high-dimensional image dataset that consists of identification (ID) cards and driving licenses. Additionally, we introduce the Siamese-based MIA that shows similar or better performance than the state-of-the-art approaches and suggest training a shadow model with autoencoder-based reconstructed images. Our major findings show that the performance of MIA techniques against too many features may be degraded; the MIA configuration or a sample's properties can impact the accuracy of membership inference on members and non-members.

**Keywords:** Membership Inference Attack · Machine Learning.

## 1 Introduction

Today, the advancements in artificial intelligence technologies lead to the wide adoption of Machine Learning as a Service (MLaaS) across various sectors and services. (*e.g.*, ChatGPT [25], Claude [2], DALL-E [26]). While MLaaS stores a model in the cloud [1,22] and processes cloud and processes the vast amounts of data remotely every day, the pervasive utilities inevitably expose a severe threat to security and privacy such as information leaks. One such concern is a membership inference attack [33] (hereinafter dubbed MIA) where an adversary attempts to determine the presence of a data sample in a training dataset while building a

* Corresponding author

model. A successful inference can accidentally reveal sensitive information such as a patient's medical record. Meanwhile, from a non-adversarial perspective, an MIA technique can be utilized to evaluate the effectiveness of machine unlearning [4] that aims to forget a particular instance or a class (*e.g.*, data deletion request to comply with regulations such as General Data Protection Regulation (GDPR) [21] and California Consumer Privacy Act (CCPA) [35].

Recent advances in MIA introduce varying techniques depending on the attacker's knowledge; black-box based [31,11,7,36,34,38,5,23,18,40,19,30,8] and white-box based [24,17] MIAs. The former setting requires a strong assumption that the adversary is aware of a victim model's architecture, parameters, and dataset, while the latter assumes that part of that information has been known. However, prior MIA approaches [31,11,36,34,38,5,18,40,19,30,8,24] evaluate their effectiveness with a standard (well-known) and (relatively simple) benchmark datasets such as MNIST [9], CIFAR-10 [15], and CIFAR-100 [15]. Besides, each MIA technique has different assumptions, rendering the potential applicability on a (high dimensional and complex) real-world dataset questionable.

In this work, we provide an in-depth empirical study of black-box based MIAs including six metric-based black-box MIAs and three classifier-based black-box MIAs; the effectiveness of different MIA approaches under the same setting with both traditional benchmark (*e.g.*, CIFAR-10 [15]) and real-world datasets (KID34K [27]: identification cards and driving licenses). We introduce a Siamese-based MIA that adopts a prediction vector generated by a shadow model alone, where the model learns the distances from sample pairs (*i.e.,* decreasing the distance between same-label-pair samples and increasing the distance between different-label-pair samples). Notably, our Siamese-based MIA achieved high performance (*i.e.,* 70% AUC) with CIFAR-10 among other state-of-the-art techniques, however, it did not work with KID34K. Additionally, we conduct thorough analyses in varying settings: training a shadow model with reconstructed images with an Autoencoder [3] so that the model can learn essential features, and splitting KID34K in a way that better discloses membership.

In summary, our major findings indicate that ① the performance of existing MIA approaches may not be persistent in a real-world dataset that contains too many features, ② the accuracy of membership inference on members and non-members can vary depending on the configuration, and ③ a sample's properties can significantly impact on the success of an MIA.

The following summarizes our contributions:

- To the best of our knowledge, we first apply an MIA to a (high dimensional) real-world dataset, demonstrating the effectiveness of MIA.
- We conduct varying experiments on different black-box MIA techniques under the same setting, including six metric-based and three classifier-based MIAs.
- We propose a Siamese-based MIA that learns the distance difference between member and non-member samples, being capable of similar membership inference performance even with limited information.
- We empirically show that reconstructed images by an autoencoder can potentially assist in training a shadow model for better MIA performance.

– We perform an in-depth analysis with different configurations: splitting a training dataset, confusion matrices, and visualizations.

## 2 Related Work

We survey varying MIAs, classifying them primarily into white-box based [16,17,24] and black-box based MIAs. The latter can be fallen into two types depending on the model construction: a metric-based attack [31,11,7,36,34,38,5,23] (generating a shadow model) and a classifier-based [18,40,19,30,8] attack (generating both shadow and classification models). Besides, we explore defense strategies [6,24,13,29,14] against MIAs.

**White-box based MIAs:** A white-box MIA assumes that the attacker has complete knowledge of the internal parameters of a victim model and significant information about the training dataset. As one may expect, such supplementary information enhances overall attack performance [24] by utilizing the intermediate computations of the victim model as inputs to build an MIA attack model. However, assuming that the adversary is aware of a significant portion of the private training dataset is unrealistic. Later, Liu et al. relaxed such an assumption on a training dataset [17]. Note that our work focuses on black box-based MIAs with a realistic setting.

**Black-box based MIAs:** A black-box MIA assumes that the attacker has query access to a victim model with the knowledge of its structure and the distribution of its training dataset. A classifier-based MIA [18,40,19,30,8] infers membership during attack model training from the outputs of a shadow model. TrajectoryMIA [18] leverages differences in the loss trajectory of a victim model during the training and testing stages. Note that knowledge distillation has been employed by training a shadow model because a black-box setting does not allow the attacker to directly obtain the victim model's loss trajectory. Meanwhile, a metric-based MIA [31,11,7,36,34,38,5,23] uses specific metrics derived from the model's outputs to determine membership. The intuition behind this technique is that the victim model has been overfitted. In ensemble models like EnsembleMIA [31], the performance improves as the number of models used increases, but privacy decreases. The fused method of ensemble models enables membership inference by averaging prediction confidence values, amplifying differences between training, and testing sample confidences to facilitate membership inference. MIB [11] infers membership by adding triggers to victim samples and verifying if the model has been backdoored.

**Defenses against MIAs:** In general, an overfitted model can be victimized by MIAs. In response, RELAXLOSS [6] and HEMP [24] employ regularization to mitigate the overfitting problem. For example, RELAXLOSS [6] leverages a training loss of a victim model with a threshold to prevent model overfitting: adopting a gradient descent during training when the loss exceeds the threshold or either a gradient ascent or posterior probability flattening otherwise. Similary, HAMP [24] attempts to mitigate overfitting with label smoothing. Meanwhile, DMIG [13] utilizes synthetic data (*e.g.*, reconstructed images) for training a

model to provide incomplete prediction vectors. MemGuard [14] appends a noise to prediction vectors to degrade the attack model's performance. Another direction leverages differential privacy to defend against MIAs, such as DPSGD [29].

## 3 Background

### 3.1 Membership Inference Attacks

**Problem Definition:** An MIA aims to accurately infer whether a specific sample has been part of the training dataset for a machine learning model. Formally, we define an MIA as in Equation 1: the attack model $g(\theta_\alpha, f)$ determines the membership in a trained machine learning model of $f(\theta_v, x)$ when a target sample $x$ is given. Note that 1 represents that $x$ is a member or 0 otherwise.

$$g(\theta_\alpha, f(\theta_v, x)) \rightarrow \{0, 1\} \tag{1}$$

**Impact:** On one hand, MIAs can pose a severe threat in terms of privacy when sensitive information (*e.g.*, medical records) has been trained. On the other hand, MIAs can be utilized in the field of machine unlearning to verify its effectiveness (*e.g.*, the unlearning request has been applied).

**Types and Assumptions:** Based on the knowledge of an attacker, MIAs can be predominately categorized into white-box and black-box attacks. The former setting [16,24] requires a strong assumption that an adversary owns the knowledge of a victim model's architecture, its model algorithm and parameters, and even a dataset. Some approaches [24,17] assume that an attacker has complete access to the victim model, being able to observe intermediate training information such as model parameters or gradients for further inference. Meanwhile, the latter setting relaxes the assumption that the adversary has limited knowledge (*e.g.*, model architecture [18,19,30,8,31], dataset distribution [34,38,5,23], or target samples [11]) about the victim model (*i.e.,* distinguishing a membership with queries and corresponding outputs).

### 3.2 Black-box MIAs

This work merely focuses on black-box MIAs that hold more realistic configurations than white-box ones, which can be classified mainly into metric-based and classifier-based MIAs (Hu et al. [12]).

**Metric-based MIAs:** Metric-based MIAs [31,11,7,36,34,38,5,23] involve querying a model to obtain a prediction vector. Simply put, prediction vectors from querying a model can be computed using a specific metric, followed by comparing those predictions with a pre-defined threshold. Note that metric-based MIAs may use shadow models (for accuracy) that approximate a target victim model, or may not use them (for efficiency). Hu et al. [12] introduce the following four types (each technique exploits different aspects of the model's output): ① a correctness-based attack infers membership if the prediction vector accurately predicts the label; ② a loss-based attack infers membership if the loss

of the prediction vector is smaller than the training loss, using a loss metric; ③ a confidence-based attack infers membership when the maximum prediction confidence exceeds a specified threshold; and ④ an entropy-based attack infers membership if a prediction vector's entropy is smaller than a specified threshold.

**Classifier-based MIAs:** A classifier-based MIA [18,40,19,30,8] involves shadow model(s)($M^S$) to mimic a victim model ($M^V$), leveraging it to create an attack dataset ($D^A$) labeled with either a member or a non-member. Then, an attack model ($M^A$) is trained to infer the membership of a given sample.

## 4 Empirical Study on Black-box based MIAs

### 4.1 On the Effectiveness of Black-box based MIAs

This work reproduces experiments with six metric-based MIAs and three classifier-based MIAs. Section 6.2 describes the effectiveness of previous MIAs.

**Metric-based MIAs:** First, the threshold-based attack [34] (loss-based) infers membership of a sample by querying the target model (*i.e.,* victim) to obtain prediction vectors and comparing their loss with a predefined threshold. Second, the threshold entropy-based attack [34] (entropy-based) infers membership of a sample by querying the target model to obtain prediction vectors and comparing their entropy loss with a predefined threshold. Third, the likelihood ratio attack (LiRA) [5] (loss-based) infers membership using loss values and logits obtained from querying a shadow model (whose structure is the same as a target model). The subsequent three attacks are dubbed as the descriptions from the implementation of Ye et al. [39]. Fourth, the population metric attack (loss-based) [39] is a model-dependent MIA that applies different attack thresholds for each target model, leveraging the dependency of loss thresholds on each model to enhance performance. Fifth, the shadow metric attack [33,39] (correctness-based) is a label-dependent MIA where the adversary infers membership using predicted labels obtained from querying the model. The adversary uses the most limited knowledge to perform the attack. Lastly, the reference metric attack [39] (loss-based) is a sample-dependent MIA where the adversary applies different attack thresholds for the target data samples. This attack trains multiple reference models using data samples excluding the target and evaluates losses for specific records. Similar to MIAs designed for summary statistics and graphical models, it uses reference models to compute the probability of the null hypothesis.

**Classifier-based MIAs:** First, logistic regression-based attack [34] trains a neural network with prediction confidences from a shadow model as features to infer membership. Second, SAMIA [40] utilizes a self-attention mechanism to infer a membership. In a nutshell, applying neural network pruning to both shadow and victim models enhances the model's memorization of training data, potentially improving the performance of membership inference. Third, the confidence-based neural network attack [33] adopts confidence vectors generated by a shadow model to train an attack model for membership inference.

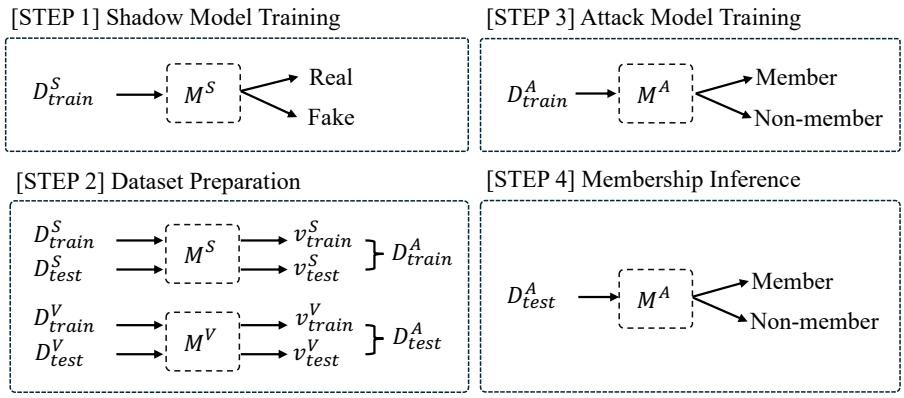



Fig. 1: Overall Siamese-based MIA workflow. An adversary generates a shadow model that mimics a target (victim) model for binary classification (Step 1). Next, the attacker prepares a dataset (Step 2) that is fed to the Siamese architecture, followed by training an attack model (Step 3). Once complete, membership can be deduced (Step 4) with the attack model. $D$ and $M$ represent a dataset and a model with their superscripts that denote a set (*i.e.,* victim, shadow, attack) and their subscripts that denote a usage (*i.e.,* training or testing).

### 4.2 Our Approach: Siamese-based MIA

We propose the Siamese network-based MIA (Figure 2) approach that learns distances between member and non-member samples.

**Assumptions:** Similar to the general black-box scenario, we assume that an adversary only has a query access to the victim model ($M^V$) with the knowledge of its architecture. Besides, the adversary owns a shadow dataset ($D^S_{train}$) that follows the same distribution as the victim model's training dataset ($D^V_{train}$).

**Overall Workflow:** Unlike the previous classifier-based MIA approach [40] that uses multiple factors (*e.g.*, confidence vector, prediction vector, one-hot label, confidence sensitivity), we use the confidence vectors generated by the shadow model ($M^S$) alone for training. This reduces the training time of the attacker model ($M^A$), saving computing resources. Figure 1 concisely illustrates the overall workflow of our Siamese-based MIA that consists of the following four stages.

**(Step 1) Shadow Model Training:** With the assumption that the distribution of a shadow dataset follows that of a victim, we divide the entire dataset into shadow and victim datasets, each consisting of 50%. Then, we split the shadow dataset into training and test sets, representing members and non-members so that the attacker can train $M^S$.

**(Step 2) Dataset Preparation:** Obtaining the predicted vectors from $M^S$ trained on $D^S_{train}$, we label those vectors for the training dataset as members and those for the test dataset as non-members. We prepare a positive set (*e.g.*,

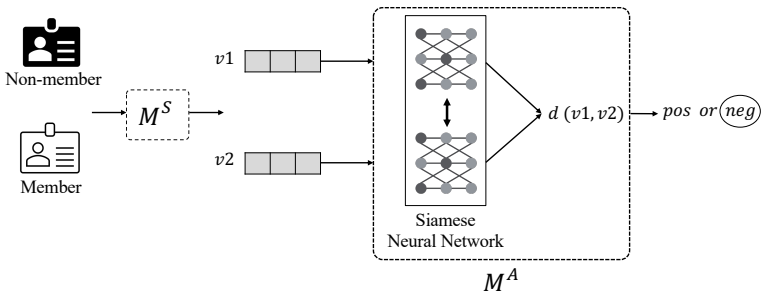



Fig. 2: Overview of our Siamese-based MIA. Taking the two predicted vectors from the member and non-member samples as input, we train an attack model so that a distance can get close for positive sets or distant for negative sets.

member to member or non-member to non-member pairs) and a negative set (*e.g.*, member to non-member pairs) for training $M^A$ with the Siamese network.

**(Step 3) Attack Model Training:** With $D^A$ from $M^S$, the adversary trains $M^A$. The Siamese network-based attack model learns distances between samples, aiming to decrease the distance between same-label-pair samples while increasing it between different-label-pair samples. The distance ($d$) between the two inputs can be computed as $d = \|p_\theta(v_1), p_\theta(v_2))\|_2$ between the two vectors where $v_1$ and $v_2$ are the output vectors of the Siamese neural network. Equation 2 represents the contrastive loss function of Siamese that minimizes the distance where $y$ represents the label (0 or 1) for the input pairs.

$$L(v_1, v_2, y) = y\|v_1 - v_2\|^2 + (1 - y) \cdot \max(0, m^2 - \|v_1 - v_2\|^2) \tag{2}$$

Note that $m$ is a hyperparameter that defines the lower bound distance between samples of different classes.

**(Step 4) Membership Inference:** Now, the adversary can infer membership with $M^A$ via $M^V$. Suppose that the adversary has non-member and unknown samples for membership inference. Then, those samples are fed into the victim model, producing predicted vectors. Finally, the attack model takes the vectors, producing the output of 1 for a member sample, or 0 otherwise.

### 4.3 Image Reconstruction with an Autoencoder

To the best of our knowledge, we first conduct varying MIAs against one of the real-world datasets, KID34K [27] (Section 6.1), rather than a common dataset like CIFAR-10 [15] or MNIST [9]. This is challenging because KID34K contains high-resolution images (*i.e.,* $512 \times 800$) with two-class classification (*i.e.,* `Real` or `Fake`). Our experimental results (Section 6.2) demonstrate that the performance of most existing approaches (including our Siamese-based MIA) has been considerably degraded.

Table 1: We split KID34K [27] into training and testing sets, each divided in half according to a victim and shadow set; 25% for $D^S_{train}$, $D^V_{train}$, $D^S_{test}$, and $D^V_{test}$. Besides, we define two datasets: $D_{rd}$ that randomly splits each dataset and $D_{sp}$ that splits by the individual user (*i.e.,* realistic assumption to check the membership of a single person). $G$, $P$, and $S$ represent a set of samples whose labels are `Genuine`, `Print`, and `Screen`. It is noted that we define a desirable set by combining sample groups: *e.g.*, $G|P$ represents a union of `Genuine` and `Print` samples.

| | | | **Train Set (Member: 50%)** | | | | | | **Test Set (Non-Member: 50%)** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **Victim** ($D^V_{train}$) | | | **Shadow** ($D^S_{train}$) | | | **Victim** ($D^V_{test}$) | | | **Shadow** ($D^S_{test}$) | | |
| **Dataset** | **Class** | **Label** | $G\|P$ | $G\|S$ | All | $G\|P$ | $G\|S$ | All | $G\|P$ | $G\|S$ | All | $G\|P$ | $G\|S$ | All |
| $D_{rd}$ | `Real` | `Genuine` | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 | 3,073 |
| | `Fake` | `Print` | 1,813 | 0 | 1,813 | 1,801 | 0 | 1,801 | 1,813 | 1,813 | 1,813 | 1,801 | 1,801 | 1,801 |
| | | `Screen` | 0 | 3,478 | 3,478 | 0 | 3,488 | 3,488 | 3,478 | 3,478 | 3,478 | 3,478 | 3,478 | 3,478 |
| | **Total** | | 4,886 | 6,551 | 8,364 | 4,874 | 6,561 | 8,362 | 8,364 | 8,364 | 8,364 | 8,352 | 8,352 | 8,352 |
| $D_{sp}$ | `Real` | `Genuine` | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 | 3,335 |
| | `Fake` | `Print` | 1,707 | 0 | 1,707 | 1,706 | 0 | 1,706 | 1,707 | 1,707 | 1,707 | 1,706 | 1,706 | 1,706 |
| | | `Screen` | 0 | 2,997 | 2,997 | 0 | 2,996 | 2,996 | 2,997 | 2,997 | 2,997 | 2,996 | 2,996 | 2,996 |
| | **Total** | | 5,042 | 6,332 | 8,039 | 5,041 | 6,331 | 8,037 | 8,039 | 8,039 | 8,039 | 8,037 | 8,037 | 8,037 |

We hypothesize that a shadow model learns excessive features (due to high dimensions). To extract significant features for decision-making, we adopt an autoencoder [3] that allows for compressing an input image to a lower dimension (*i.e.,* latent vector). Then, we reconstruct images that retain essential features for training a shadow model.

## 5 Implementation

**Victim and Shadow Models:** We adopt ResNet20 [10] and ResNet50 [10] as the victim and shadow models provided by the PyTorch [28]'s `timm` library (PyTorch Image Models) [37]. We set the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-3}$, and the training epochs to 100 for both models. Each model's training took around 4 hours with approximately $8,000$ samples. Another hyperparameter is the number of shadow models for each MIA: two for the logistic regression attack [34] and LiRA [5], five for SAMIA [40] and the confidence-based neural network attack [33], and only one for our Siamese-based MIA. It is noteworthy to mention that increasing the number of shadow models does not guarantee higher MIA performance at all times.

**Siamese-based MIA:** We developed the Siamese-based MIA framework based on the original Siamese structure with PyTorch [28]. Training the model with around 8,000 samples took about 3 hours. We set up the Adam optimizer with $\epsilon = 10^{-3}$ and the model training epochs to 100.

**Image Reconstruction with an Autoencoder:** We implemented the Autoencoder [3] architecture with PyTorch [28] for creating reconstructed images. The autoencoder training took approximately 12 hours with around 8,000 samples from our shadow dataset. We set up the Adam optimizer with $\epsilon = 10^{-6}$ and the training epochs to 200.

## 6 Evaluation

### 6.1 Experimental Setup

Our experiments were conducted on a server equipped with an Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz, 500GB RAM, and three Matrox Electronics Systems MGA G200e [Pilot] with 480GB of memory.

**Datasets:** For evaluating the effectiveness of varying black-box based MIAs, we use both a common (CIFAR-10 [15]) and a real-world (KID34K [27]) dataset. CIFAR-10 [15] is an image dataset that consists of different objects with a $3\times32\times32$ resolution, which contains 10 classes, with 6,000 samples per class, a total of 60,000 samples. KID34K [27] is an image dataset comprising identification cards and driving licenses with a $3 \times 512 \times 800$ resolution. It contains two classes: `Real` and `Fake`. The `Real` class contains samples with `genuine` labels, while the `Fake` class includes samples labeled as `print` (*i.e.,* paper-printed image) or `screen` (*i.e.,* screen-captured image). There are 10,488, 13,728, and 10,444 samples for `genuine`, `screen`, and `print`, respectively. Table 1 shows how we split KID34K into training and test datasets with each dividing in half to a victim and a shadow set.

**Evaluation Metrics:** We adopt F1 and AUC (Area Under Curve) as evaluation metrics. F1 is the harmonic mean of precision and recall. AUC is computed as the area under the ROC (Receiver Operating Characteristic) curve, which illustrates the change in TPR (True Positive Rate) with respect to FPR (False Positive Rate) variations.

**Research Questions:** We define the following three research questions (RQs) to answer the effectiveness and applicability of various MIAs.

- RQ1. How effective are black-box based MIAs (including our Siamese-based MIA) on a previous benchmark dataset (*i.e.,* CIFAR-10 [15])(Section 6.2)?
- RQ2. How well do the MIAs perform against a real-world dataset (KID34K [27]) (Section 6.3)?
- RQ3. How well the reconstructed images improve MIA performance on a real-world dataset (Section 6.4)?

### 6.2 Comparison of Different Black-box MIA Approaches (RQ1)

Using CIFAR-10, we conduct experiments with nine black-box MIA techniques as a baseline, comparing them with our Siamese-based MIA. Table 2 summarizes

Table 2: Empirical results of varying MIAs against the victim model trained on CIFAR-10. The re-evaluation of metric-based MIAs demonstrates discrepancies ($< 0.6$) with the original performance. While binary classifier-based approaches show better performance, our Siamese-based MIA ranks the highest AUC (0.7).

| **Attack Technique** | **Base Approach** | **AUC** |
|---|---|---|
| MIA Evaluation (Threshold attack) [34] | Metric-based (Loss) | 0.52 |
| MIA Evaluation (Threshold entropy attack) [34] | Metric-based (Entropy) | 0.51 |
| LiRA [5] | Metric-based (Loss) | 0.58 |
| Privacy Meter (Population metric attack) [39] | Metric-based (Loss) | 0.50 |
| Privacy Meter (Shadow metric attack) [33,39] | Metric-based (Correctness) | 0.50 |
| Privacy Meter (Reference metric attack) [39] | Metric-based (Loss) | 0.50 |
| MIA Evaluation (Logistic regression attack) [34] | Classifier-based | 0.51 |
| SAMIA [40] | Classifier-based | 0.67 |
| Confidence-based neural network attack [33] | Classifier-based | 0.64 |
| Siamese-based MIA (Ours) | Classifier-based | 0.70 |

comparison results with the CIFAR-10 dataset. Our findings show that ① metric-based MIAs tend to have discrepancies with the original performance ($< 0.6$), ② binary-classifier-based MIAs better perform than metric-based ones, and ③ our Siamese-based MIA achieves the highest AUC (0.7) with efficiency.

### 6.3 Effectiveness of Black-box MIAs on a Real-World Dataset (RQ2)

We assess varying black-box MIAs against KID34K [27], one of the real-world datasets as a victim model. Notably, we choose the two MIA techniques (*i.e.,* SAMIA [40] and confidence-based neural network [33]) that exhibit relatively higher performance. Table 3 presents a handful of interesting findings. First, the results indicate that an MIA approach against a common dataset may not be persistent in a real-world dataset. For example, the AUCs of our Siamese-based approach demonstrate slightly behind SAMIA and the confidence-based technique. Second, membership inference accuracy on a member ($A_m$) and a non-member ($A_n$) sample can largely vary depending on a configuration and a dataset. Even with the same approach (*e.g.*, SAMIA), the accuracy on deducing a non-member ($A_n$) is higher than a member ($A_m$) in a random selection setting, however, the other way around in a user split setting. Third, it is possible that the success of an MIA may vary depending on the sample's property (*e.g.*, randomly selected sample VS. member-oriented sample) as well as an MIA technique.

### 6.4 Effectiveness of Autoencoder-reconstructed Images (RQ3)

Table 3 indicates that attack models using reconstructed images improves (or is on par with) attack success rates (both F1 score and AUC) where overall

Table 3: We conduct a variety of MIA experiments on a real-world dataset. The baselines are SAMIA and the confidence-based neural network, which demonstrates high performance. We prepare a shadow dataset with different configurations to demonstrate that overall MIA performance can be affected by i) a data sample (by splitting it into $D_{rd}$ and $D_{sp}$), ii) reconstructed images, and iii) MIA approach. In our experiment, combining SAMIA with reconstructed images of $D_{sp}$ using an autoencoder shows the best performance. A, P, and R represent accuracy, precision, and recall while $A_m$ and $A_n$ denote the accuracy of a member and a non-member. (*) means our Siamese-based MIA. A bold number represents better performance between original and reconstructed corpus from $D_{sp}$.

| **Corpus** | **Selection** | $D_{train}$ | **Method** | $A_m$ | $A_n$ | **A** | **P** | **R** | **F1** | **AUC** |
|---|---|---|---|---|---|---|---|---|---|---|
| Original | $D_{rd}$ | $G\|P$ | SAMIA | 0.17 | 0.83 | 0.50 | 0.50 | 0.50 | 0.45 | 0.50 |
| | $D_{rd}$ | $G\|S$ | SAMIA | 0.30 | 0.70 | 0.50 | 0.50 | 0.50 | 0.48 | 0.50 |
| | $D_{rd}$ | *All* | SAMIA | 0.21 | 0.79 | 0.50 | 0.50 | 0.50 | 0.45 | 0.50 |
| | $D_{sp}$ | $G\|P$ | SAMIA | 0.99 | 0.25 | 0.62 | 0.77 | 0.62 | 0.56 | 0.62 |
| | $D_{sp}$ | $G\|S$ | SAMIA | 0.61 | 0.45 | 0.53 | 0.55 | 0.53 | 0.53 | 0.53 |
| | $D_{sp}$ | *All* | SAMIA | 0.73 | 0.35 | **0.54** | 0.56 | 0.54 | 0.52 | **0.54** |
| | $D_{rd}$ | $G\|P$ | Confidence-based | 0.18 | 0.83 | 0.50 | 0.51 | 0.50 | 0.45 | 0.50 |
| | $D_{rd}$ | $G\|S$ | Confidence-based | 0.75 | 0.25 | 0.50 | 0.50 | 0.50 | 0.47 | 0.50 |
| | $D_{rd}$ | *All* | Confidence-based | 0.20 | 0.81 | 0.50 | 0.50 | 0.50 | 0.45 | 0.50 |
| | $D_{sp}$ | $G\|P$ | Confidence-based | 0.35 | 0.99 | **0.67** | 0.79 | 0.67 | **0.63** | **0.67** |
| | $D_{sp}$ | $G\|S$ | Confidence-based | 0.53 | 0.56 | 0.54 | 0.56 | 0.54 | 0.54 | 0.54 |
| | $D_{sp}$ | *All* | Confidence-based | 0.70 | 0.38 | **0.54** | 0.56 | 0.54 | 0.53 | **0.54** |
| | $D_{rd}$ | $G\|P$ | Siamese-based* | 0.14 | 0.85 | 0.50 | 0.50 | 0.50 | 0.42 | 0.50 |
| | $D_{rd}$ | $G\|S$ | Siamese-based* | 0.15 | 0.85 | 0.50 | 0.50 | 0.50 | 0.43 | 0.50 |
| | $D_{rd}$ | *All* | Siamese-based* | 1.00 | 0.00 | 0.50 | 0.25 | 0.50 | 0.33 | 0.50 |
| | $D_{sp}$ | $G\|P$ | Siamese-based* | 0.19 | 0.83 | 0.51 | 0.53 | 0.51 | 0.45 | 0.51 |
| | $D_{sp}$ | $G\|S$ | Siamese-based* | 0.13 | 0.82 | 0.48 | 0.47 | 0.48 | 0.41 | 0.47 |
| | $D_{sp}$ | *All* | Siamese-based* | 0.17 | 0.82 | **0.50** | 0.51 | 0.50 | 0.44 | **0.50** |
| Reconstructed | $D_{sp}$ | $G\|P$ | SAMIA | 0.40 | 0.93 | **0.66** | 0.74 | 0.66 | **0.64** | **0.66** |
| | $D_{sp}$ | $G\|S$ | SAMIA | 0.67 | 0.41 | **0.54** | 0.56 | 0.54 | 0.53 | **0.54** |
| | $D_{sp}$ | *All* | SAMIA | 0.56 | 0.50 | 0.53 | 0.54 | 0.53 | **0.53** | 0.53 |
| | $D_{sp}$ | $G\|P$ | Confidence-based | 0.36 | 0.95 | 0.66 | 0.75 | 0.66 | 0.62 | 0.66 |
| | $D_{sp}$ | $G\|S$ | Confidence-based | 0.60 | 0.49 | 0.54 | 0.56 | 0.54 | 0.54 | 0.54 |
| | $D_{sp}$ | *All* | Confidence-based | 0.63 | 0.44 | 0.53 | 0.55 | 0.53 | 0.53 | 0.53 |
| | $D_{sp}$ | $G\|P$ | Siamese-based* | 0.54 | 0.50 | **0.52** | 0.54 | 0.52 | **0.51** | **0.52** |
| | $D_{sp}$ | $G\|S$ | Siamese-based* | 0.16 | 0.80 | 0.48 | 0.48 | 0.48 | **0.42** | **0.48** |
| | $D_{sp}$ | *All* | Siamese-based* | 0.22 | 0.75 | 0.49 | 0.50 | 0.49 | **0.45** | 0.49 |

SAMIA with the autoencoder-generated samples ranks the highest (AUC of 0.66). For instance, our proposed Siamese-based MIA showed a 5.8% increase in F1 in the $G|P$ setting with $D_{sp}$. For SAMIA [40], an 8.1% increase in F1 was achieved in the same configuration, observing overall increases in most other settings. However, in the case of confidence-based neural network attack [33], AUC increase was limited (around 0.3 %) in the $G|P$ configuration with $D_{sp}$. Additionally, our experiment on the effectiveness of a different size of latent vectors (*e.g.*, 128 vs. 512) shows limited impact on the performance of a shadow

Table 4: Experimental results of shadow models' performance using the whole $D_{train}$ (*i.e., All*) according to a dimension ($d$) in a latent space. We adopt $d = 128$ because of the little gap in performance.

| **Method** | $d$ | **A** | **P** | **R** | **F1** | **AUC** |
|---|---|---|---|---|---|---|
| SAMIA | 128 | 0.53 | 0.54 | 0.53 | 0.53 | 0.53 |
| | 512 | 0.53 | 0.54 | 0.53 | 0.53 | 0.53 |
| Confidence-based | 128 | 0.53 | 0.55 | 0.53 | 0.53 | 0.53 |
| | 512 | 0.53 | 0.55 | 0.53 | 0.53 | 0.53 |
| Siamese-based* | 128 | 0.49 | 0.50 | 0.49 | 0.45 | 0.49 |
| | 512 | 0.50 | 0.51 | 0.50 | 0.48 | 0.50 |

(a) $D_{rd}$ (TP) (b) $D_{rd}$ (TN)

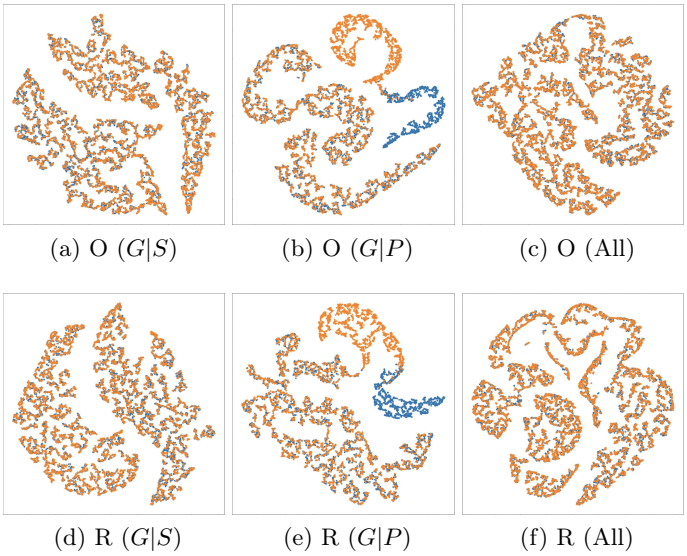
(c) $D_{sp}$ (TP) (d) $D_{sp}$ (TN)

Fig. 3: Grad-CAM [32] results of a shadow model that is trained on the randomly picking dataset ($D_{rd}$) and the individually splitting dataset ($D_{sp}$). The model sees more features in $D_{sp}$ for true positive samples (*e.g.*, `Real` $\rightarrow$ `Real`) than those in $D_{rd}$. Meanwhile, the model recognizes relatively plentiful features in both $D_{rd}$ and $D_{sp}$ for true negative samples (*e.g.*, `Fake` $\rightarrow$ `Fake`). TP and TN denote a true positive and negative.

model. Note that we use $d = 128$. Table 4 summarizes the experimental results of three shadow models: SAMIA, confidence-based, and Siamese-based MIAs.

### 6.5 In-depth Analysis of Black-box MIAs on a Real-World Dataset

This section delves into in-depth analysis of the results of our MIA experiments in Table 3. According to the description of the KID34K [27] dataset, it contains the images of identification cards from 37 users and driver licenses from 45 users,



(a) O ($G|S$) (b) O ($G|P$) (c) O (All)

(d) R ($G|S$) (e) R ($G|P$) (f) R (All)

Fig. 4: t-SNE [20] results of SAMIA with $D_{sp}$. This visualization illustrates the reason why the shadow model trained with $D_{train}$ that consists of $G|P$ (*e.g.*, relatively clear boundaries for decision-making) achieves the highest F1 and AUC (Table 3). O and R denote the original and reconstructed images, respectively.

with a total of 46 users. Because the random selection of a dataset ($D_{rd}$) lowers the performance of a shadow model, we split it based on an individual sample ($D_{sp}$). We visualize the results of SAMIA with $D_{sp}$ using t-SNE in Figure 4. In the $G|P$ cases (*e.g.*, datasets picking `genuine` and `print` labels) of both original and reconstructed images, the model recognizes clear boundaries for membership decisions. Additionally, Figure 3 illustrates Grad-CAM [32], showing that the shadow model with $D_{sp}$ identifies more features than that with $D_{rd}$ for true positive samples (Figure 3 (a) and Figure 3 (c)).

## 7 Discussion & Limitations

**Threats to Validity:** We empirically demonstrate that MIA results may vary depending on the configuration, the distribution of a dataset, the number of classes, and the characteristics of a sample as well as an approach. We believe that a single counterexample can be allusive to convey that MIAs against arbitrary datasets can be ineffective. However, the results with other real-world datasets may be possibly inconsistent with ours: *i.e.,* the dataset in our experiment is not fully representative with different features (financial or health data). As a final note, this work focuses solely on black-box MIAs under relaxed assumptions; hence the results with white-box MIAs may be different.

**Usage of Reconstructed Images for MIAs:** Although using autoencoder-generated images for training a shadow model enhances the success rate of MIAs, we observe the degradation of the shadow model's performance (around 16%). While an adversary can adopt this strategy for reducing dimensions, optimizing the size of a latent vector that encompasses essential features is an open problem.

**Machine Learning Bill of Materials (ML-BOM) and MIAs:** As the demand for reliability in machine learning services increases, a comprehensive inventory that documents the whole process of model creation, deployment, and maintenance is needed. This request inspires the Machine Learning Bill of Materials (ML-BOM), ensuring transparency, accountability, traceability, and compliance. We expect that the assumption of white-box based MIAs become feasible upon the broad adoption of ML-BOM in the near future.

**Trade-offs between Membership Inference Attacks and Defenses:** In general, defending against MIAs often involves trade-offs between model performance (*i.e.,* usability, accuracy) and security (*i.e.,* privacy, robustness). For instance, a model with high-performance could be susceptible and targetable to MIAs; however, security mitigations such as differential privacy, model distillation, or adopting multiparty computation could be prone to reduce performance with additional overheads.

## 8 Conclusion

Lately, the prevalent utilities of machine learning-based applications have raised concerns about security and privacy issues. An MIA that determines the presence of a sample in a training set can pose a severe threat, when sensitive information such as one's medical record, driver license, passport, or identification card information is inadvertently revealed. In this work, we conduct an empirical study focusing on black-box based MIAs against a real-world dataset, KID34K dataset, which contains driver license and identification card information. Through extensive experimental evaluation, our findings reveal that the performance of existing MIAs could be degraded and impacted by their settings or a sample's properties.

## Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work was partially supported by three grants from Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korean government (MSIT; Ministry of Science and ICT): No. 2022-0-00688; No. RS-2024-00337414; and No. 2022-0-01199. Additional support was provided by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education of the Government of South Korea (Grant No. NRF-2022R1F1A1074373). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

## References


1. Amazon: Cloud computing services - amazon web services. https://aws.amazon.com/ko/ (2024)
2. Anthropic: Generate code with claude. https://claude.ai (2024)
3. Baldi, P.: Autoencoders, unsupervised learning, and deep architectures. In: ICML workshop on unsupervised and transfer learning (2012)
4. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D., Papernot, N.: Machine unlearning. In: IEEE Symposium on Security and Privacy (2021)
5. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: IEEE Symposium on Security and Privacy (2022)
6. Chen, D., Yu, N., Fritz, M.: Relaxloss: Defending membership inference attacks without losing utility. arXiv preprint arXiv:2207.05801 (2022)
7. Chen, Y., Shen, C., Shen, Y., Wang, C., Zhang, Y.: Amplifying membership exposure via data poisoning. Advances in Neural Information Processing Systems (2022)
8. Choquette-Choo, C.A., Tramer, F., Carlini, N., Papernot, N.: Label-only membership inference attacks. In: International Conference on Machine Learning (2021)
9. Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine (2012)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
11. Hu, H., Salcic, Z., Dobbie, G., Chen, J., Sun, L., Zhang, X.: Membership inference via backdooring. arXiv preprint arXiv:2206.04823 (2022)
12. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys (2022)
13. Hu, L., Li, J., Lin, G., Peng, S., Zhang, Z., Zhang, Y., Dong, C.: Defending against membership inference attacks with high utility by gan. IEEE Transactions on Dependable and Secure Computing (2022)
14. Jia, J., Salem, A., Backes, M., Zhang, Y., Gong, N.Z.: Memguard: Defending against black-box membership inference attacks via adversarial examples. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security (2019)
15. Krizhevsky, A.: The cifar-10/100 dataset. https://www.cs.toronto.edu/~kriz/cifar.html (2024)
16. Leino, K., Fredrikson, M.: Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In: USENIX Security Symposium (2020)
17. Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., Tang, J.: Trustworthy ai: A computational perspective. ACM Transactions on Intelligent Systems and Technology (2022)
18. Liu, Y., Zhao, Z., Backes, M., Zhang, Y.: Membership inference attacks by exploiting loss trajectory. In: ACM SIGSAC Conference on Computer and Communications Security (2022)
19. Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., De Cristofaro, E., Fritz, M., Zhang, Y.: {ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models. In: USENIX Security Symposium (2022)

20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (2008)
21. Mantelero, A.: The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. Computer Law & Security Review (2013)
22. Microsoft: Microsoft azure: Cloud computing services. https://azure.microsoft.com/en-us (2024)
23. Murakonda, S.K., Shokri, R.: Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. arXiv 2007.09339 (2020)
24. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: IEEE Symposium on Security and Privacy. pp. 739–753 (2019)
25. OpenAI: Chatgpt. https://openai.com/index/chatgpt/ (2022)
26. OpenAI: Dall·e. https://openai.com/index/dall-e/ (2024)
27. Park, E.J., Back, S.Y., Kim, J., Woo, S.S.: Kid34k: A dataset for online identity card fraud detection. In: ACM International Conference on Information and Knowledge Management (2023)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (2019)
29. Rahimian, S., Orekondy, T., Fritz, M.: Differential privacy defenses and sampling attacks for membership inference. In: ACM workshop on artificial intelligence and security (2021)
30. Rezaei, S., Liu, X.: On the difficulty of membership inference attacks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
31. Rezaei, S., Shafiq, Z., Liu, X.: Accuracy-privacy trade-off in deep ensemble: A membership inference perspective. In: IEEE Symposium on Security and Privacy (2023)
32. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (2017)
33. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy (2017)
34. Song, L., Mittal, P.: Systematic evaluation of privacy risks of machine learning models. In: USENIX Security Symposium (2021)
35. de la Torre, L.: A guide to the california consumer privacy act of 2018. Available at SSRN 3275571 (2018)
36. Wen, Y., Bansal, A., Kazemi, H., Borgnia, E., Goldblum, M., Geiping, J., Goldstein, T.: Canary in a coalmine: Better membership inference with ensembled adversarial queries. arXiv preprint arXiv:2210.10750 (2022)
37. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019)
38. Ye, J., Maddi, A., Murakonda, S.K., Bindschaedler, V., Shokri, R.: Enhanced membership inference attacks against machine learning models. In: ACM SIGSAC Conference on Computer and Communications Security. pp. 3093–3106 (2022)
39. Ye, J., Maddi, A., Murakonda, S.K., Shokri, R.: Privacy auditing of machine learning using membership inference attacks. arXiv preprint arXiv:2108.04417 (2021)
40. Yuan, X., Zhang, L.: Membership inference attacks and defenses in neural network pruning. In: USENIX Security Symposium (2022)